Assignment IV : Classification - Random Forest, Adaboost and KNN

March 19, 2025

ID5002W: Industrial AI Laboratory

Total Marks: 80

Instructions

1. Assignment shall be submitted before the due date. Late submissions will not be entertained. If you cannot submit the assignment due to some reasons, please contact Dr. Tirthankar by email.

2. All the assignments must be the student's own work. The students are encouraged to discuss or consult friends or classmates. However, they have to submit their own work. Any malpractice will be reported to the authorities and action will be taken as per the IIT Madras rules.

3. If you find the solution in the book or article or on the website, please indicate the reference in the solution

4. You are expected to submit "*.py" file instead of notebook. Follow the steps mentioned below:

   (a) Use the datasets provided to you for solving the assignment problems classification.

   (b) Create one notebook for Problem 1 and another notebook for Problem 2.

   (c) If you have used

   ```
   df = pd.read_csv("assignment4.csv")
   ```

   where "assignment4.csv" is the file provided to you, replace the code with the following before submitting the assignment.
   For Problem 1,

   ```
   dataset = "./data/problem1/assignment4.csv"
   df = pd.read_csv(dataset)
   ```

   For Problem 2,

   ```
   dataset = "./data/problem2/assignment4.csv"
   df = pd.read_csv(dataset)
   ```

   (d) Now, download each notebook as a "*.py" file and submit both the "*.py" files.

5. A short report in pdf format along with the code should be submitted containing results and analysis as asked in the assignment.

6. Please note:

   - Code should execute without any error.
   - The code should be clean with readable comments.
   - The output of the code should be in a relevant format so that it can be understood by an evaluator.

7. Grading Policy:

   - 50% code correctness.
   - 10% code readability and comments.
   - 40% on report and analysis.

Problem 1

Develop, optimize, and evaluate classification models using Random forest, Adaboost, and KNN classifiers. The dataset given ("assignment4.csv") is a complex synthetic dataset that has high dimensionality and noise. It contains 5000 samples and 50 features, along with a target column that indicates the class labels 0, 1 and 2. The goal is to perform the following tasks with an 80:20 train: test ratio and a seed value of 40 wherever relevant.

(a) Load the dataset, check for missing values (handle them if present), and scale the data (Standard scaler). [**Marks:5**]

(b) Build the Random forest, Adaboost (with decision tree as base model), and KNN classifiers and perform hyperparameter tuning for each classifier[**Marks:15**]

(c) Evaluate the best models on the test set and record the performance metrics (Accuracy, precision, recall, F1 score) [**Marks:10**]

(d) Use k-fold cross-validation with k=5 and summarize the performance of all three models. [**Marks:10**]

(e) Plot the confusion matrix for each classifier [**Marks:5**]

(f) Comment on the suitability of the models developed for the given dataset [**Marks:5**]

Problem 2

Perform feature selection for Random forest and Adaboost models and repeat steps (c) and (e) from problem 1 and comment on the model's performance. [**Marks:30**]